Master's Thesis

Augmenting Information Propagation Models with Graph Neural Networks

Wonjun Hwang

Department of Computer Science and Engineering

Ulsan National Institute of Science and Technology

2021

Augmenting Information Propagation Models with Graph Neural Networks

Wonjun Hwang

Department of Computer Science and Engineering

Ulsan National Institute of Science and Technology

Augmenting Information Propagation Models with Graph Neural Networks

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Wonjun Hwang

12/18/2021
Approved by

_____

Advisor
Youngbin Im

Augmenting Information Propagation Models with Graph Neural Networks

Wonjun Hwang

This certifies that the thesis/dissertation of Wonjun Hwang is approved.

12/18/2021

Signature

_____

Advisor: Youngbin Im

Signature

_____

typed name: Kyunghan Lee

Signature

_____

typed name: Hyoil Kim

Signature

Abstract

Conventional epidemic models are limited in their ability to capture the dynamics of real world epidemics in a sense that they either place restrictions on the models such as their topology and contact process for mathematical tractability, or focus only on the average global behavior, which lacks details for further analysis. We propose a novel modeling approach that augments the conventional epidemic models using Graph Neural Networks to improve their expressive power while preserving the useful mathematical structures. Simulation results show that our proposed model can predict spread times in both node-level and network-wide perspectives with high accuracy having median relative errors below 15% for a wide range of scenarios.

Contents

List of Figures

Introduction and Related Works

Prior works on epidemic models often place restrictions on the distribution of the contact patterns between the restrictions on the distribution of the contact patterns between the nodes and the topology of the network for mathematical tractability.

For example, prior works on epidemic modeling such as [8],[10] fails to go beyond average behavior and even the attempts to give relatively precise predictions [1] places Markovian assumptions on inter-contact processes along with topological restrictions.

Such restrictions prevent epidemic models from accurately modeling the real-world epidemics. For example, if the infection is quarantined and through the breach in the quarantine system the infection spreads, we would want contact patterns to follow Weibull distribution which is often used for the reliability engineering. Or the topology of the network under our analysis might be different from the models' restrictions. A fully data driven approach might be considered but acquiring high quality datasets on epidemics is very difficult due to many practical concerns including the privacy issue and the technical issue from knowing who have been infected and cured especially when we are facing a novel epidemic like CoVID-19.

To tackle these challenges, in this paper we propose a new hybrid modeling approach: augmenting the epidemic models with Graph Neural Networks (GNNs) which is the type of neural network that learns the dependencies between two nodes in a graph to improve their expressive power while preserving the useful mathematical structures.

We take the following steps to detail the efficacy our approach:

1) We revisit the conventional approaches in Epidemic modeling.
2) We review the recent advances in Graph Neural Networks, the deep learning architecture that specialized in handling graph data.
3) We detail the methodology of augmenting Epidemic models with Graph Neural Network.
4) We finally show that our augmentation approach yields promising results is various scenarios.

Our proposed model is powerful in that it can be applied to a wide range of contact patterns from light-tailed to heavy-tailed distributions under a general network topology and is flexible enough to

accommodate possibly heterogeneous nodes with different infection rates. Based on our augmented epidemic model, we analyze spread times from node-to-network level perspectives by introducing local completion time and global completion time for a certain level of spread.

We show the efficacy of our model through extensive simulation studies on synthetic and real-world heterogeneous networks. We find that our approach gives highly accurate predictions for networks with diverse contact patterns and topology freeing the epidemic models of their restrictions.

Epidemic Models Primer

We consider a network represented by a graph $G = (V, E)$. Each node $v \in V$ can be in state of being susceptible (S) or Infected (I) during a spreading process. The spreading process starts with initially infected nodes $i \in V$ which is called the seeds. Each seed i attempts to spread the infection to adjacent susceptible nodes j such that $(i, j) \in E$ when they contact. This attempt by infected node I succeeds with probability ϕ_i which is called the infectivity of node i. On successful reception of the infection, susceptible node j actually become infected and turns into an infected node with probability γ_j which is called the susceptibility of node j. Once a node becomes infected it behaves identically to the seed. The rest of the spreading process is then carried out in a similar manner by infected nodes until all reachable nodes in the network become infected.

In this paper, we analyze the local completion time $L(X, v)$ and global completion time $G(Y)$ as performance metrics on spread time, which we define as follows.

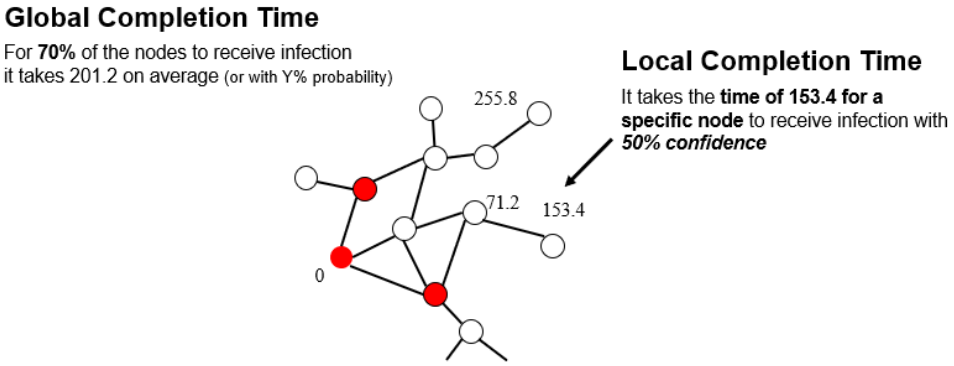

Fig. 1 Global and Local Completion Time of a Network

Definition 1. (Local Completion Time). We define $L(x, v)$ to be the time needed to guarantee the infection of a node $v \in V$ with probability X:

$$L(X, v) := \min\{ P(I_v(t) = 1) > X\}$$

where $I_v(t)$ is the indicator variable defined for node v as $I_v(t) = 1$ if it is infected at time t or $I_v(t) = 0$ otherwise.

The local completion time is useful for answering the questions on how long it takes for an individual node to become infected with X percent guarantee, or, given the current time, how likely it is for a specific node to be in infection state. It is sometimes critical to test people whether they are infected or not with limited resources in a timely manner.

A plausible solution is to utilize our resource on testing the people v with $:(X, v) < current\ time$ since those are the one who are more likely to be in a infection state.

Definition 2. (Global Completion Time). We define $G(Y)$ to be the expected time taken for Y fraction of the total nodes in the network to become infected:

$$G(Y) \coloneqq E[\min\left\{t \geq 0 \middle| \frac{|\{v \in V | I_v(t) = 1\}|}{|V|}\right\} > Y]]$$

Where $|A|$ denotes the size of a set A.

The global completion time informs us how long it will take for Y fraction of the nodes in the entire network to become infected. We can exploit $G(Y)$ in a network-wide application scenario such as seed selection problems for expediting the spread over the entire network or estimation of how much of the total population is infected at a given time.

Prior analysis of information spreading places restrictions on inter-contact time between nodes, network topology or property of information being propagated etc. Similar works by [1] has restrictions on inter-contact time with exponential distribution and its topology is fixed to fully connected topology which is the specific case where infection time can be shown to follow exponential distribution under the assumption that inter-contact time is exponential.

Those models cannot be applied to many real world scenarios where the underlying topology may be arbitrary with arbitrary intercontact times between nodes with heterogeneous properties. For example in the cases such as computer virus spread in the presence of anti-virus software screening the packets transferred between nodes inter contact time should be better modeled as Weibull distribution as in reliability engineering(failure analysis) since the spread occurs only with the failure of the anti virus software to screen the packets sent by infected computer.

Also, as in recent cases of COVID-19 the virus causes mutation changing its properties and its behavior. Due to the limitations of conventional mathematical analysis we should augment our analysis using data driven methods.

We specify the details of our deep learning model we use to predict Completion Times of epidemic models. Our deep learning model is based on Graph Neural Network(GNN), deep learning architectures specialized in performing deep learning on graph structured data [5][6].

GNN overview

Let $G = (V, E)$ be directed or undirected graph with nodes $v \in V$ and edges $(v, u) \in E$. Inputs to GNNs consist of input node features $X \in R^{nd}$ and adjacency information $A \in R^{n^2}$ where n is the number of nodes in the graph and d is the dimension of node level input features.

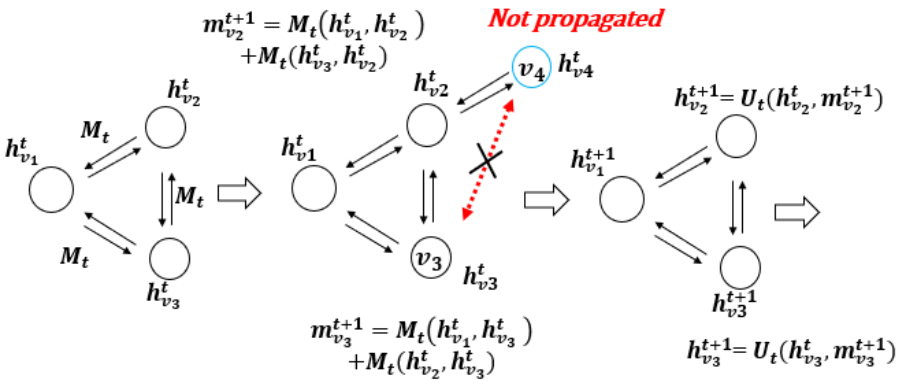

Fig. 2 Message Passing in Graph Neural Networks

Graph Neural Networks often consist of T iterative steps of Message Passing Phase followed by Readout Phase[3][4]. Single progression of GNN layers as in Fig. 2 is equivalent to one or multiple Message Passing Phase. Message Passing Phase take the topology of the graph data into account by iteratively propagating and aggregating the node features only between adjacent nodes according to adjacency information A.

The hidden representations h_v^t of $v \in V$ at step $t \in T$ of Message Passing are propagated as messages m_t between adjacent nodes by the message function M_t, Each node then aggregate the messages arrived from neighboring nodes and updates its hidden representation using update function U_t. Single iteration of Message Passing Phase could be formally described as:

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}),$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1})$$

Where node $N(v)$ is the set of nodes adjacent to nodes $v \in V$ and h_v^0 is the initial input the node v in $X = [h_v^0]$. Finally the Readout Phase computes the outputs of the whole graph with readout function R.

$$\hat{y} = R(\{h_v^k | v \in V, k = 0,1, \dots, T\})$$

All functions in Message Passing phase and Readout Phase can be any trainable functions.

The core concept of Graph Neural Networks is that it encodes the structure and topology of a graph data by carrying out the message passing phase only between adjacent nodes. As in fig2 there is no direct link between v_3 and v_4 so the hidden representations h_{v3}, h_{v4} cannot be exchanged in a single Message Passing. By propagating the node features only through neighboring nodes GNNs learn the relationships between the nodes regarding the structure of the graph.

Graph Attention Networks

Recently proposed Graph Attention Networks (GAT) [2] enhances GNN architectures by using attention mechanism which computes the level of importance (attention coefficient) that should be placed on hidden representation h_j of node j when updating hidden representation h_i of node i in the Message Passing Phase. Attention coefficient e_i is calculated using the attention function $a: R^{F\prime} \times R^{F'}$

$$e_{ij} = a(Wh_i, Wh_j)$$

Where W is a trainable weight matrix that can be learned for sufficient expressive power. Each GAT layer performs masked attention which means that attention coefficients e_{ij} are computed only for nodes $j \in N(i)$ to take the topological structure of a graph into account according to $A \in R^N \times R^N$. Attention coefficients are then normalized across all $j \in N(v)$ using softmax function:

$$\alpha_{ij} = softmax(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})}$$

As in the original paper our attention function a is a single later neural network which we parametrize by a weight vector $a^{\rightarrow} \in R^{2F'}$ followed by ReLU function for nonlinearity σ. Overall normalized

attention coefficient could be fully written as:

$$\alpha_{ij} = \frac{\exp\left(\sigma(\alpha^{\rightarrow T}[Wh_i||Wh_j])\right)}{\sum_{k\in N(i)} \exp\left(\sigma(a^T[Wh_i||Wh_k])\right)}.$$

Where || and T are concatenation and transposition operator. Normalized attention coefficients α_{ij} assigns the importance on node $j's$ hidden representation when updating $i's$ hidden representation through update function:

$$h_i^{\rightarrow} = \sigma\left(\sum_{k\in N(i)} \alpha W h_j^{\rightarrow}\right)$$

For stable training of the GAT as suggested in the original paper we employ multi-headed attention mechanism where we concatenate K independent attention mechanism:

$$h_i^{\rightarrow} = ||_k \sigma\left(\sum_{k\in N(i)} \alpha_{ij}^k W^k h_j^{\rightarrow}\right)$$

α^k are normalized attention coefficients of the kth attention head a^k and W^k are the corresponding weight matrices. In the final GAT layer we average the attention heads by:

$$h_i^{\rightarrow} = \sigma\left(\sum_k \frac{1}{K} \sum_{k\in N(i)} \alpha_{ij}^k W^k h_j\right)$$

Readout Phase

In our readout phase [3][4].of the model the model could give either node level prediction $\widehat{y_v}$ where we get output for every node in the graph or graph level prediction $\widehat{y_G}$ where readout phase has additional step that aggregates all the node level features h_v^t in the final layer to give predictions for the whole graph. We place additional dense layers at the end of GAT layers for graph level prediction.

$$\widehat{y_v} = GAT_V(h^{T-1})$$

$$\widehat{y_G} = Dense(GAT_T)$$

Graph Attention Networks are capable of inductive learning which means that they can give predictions for graphs that are completely unseen during training. This enables us to use GAT for giving predictions to completely new networks if the node features are preserved in a consistent manner.

Augmenting Epidemic Models with GNN

In this section, we detail the process of augmenting the epidemic models using GNN. Inputs to our GNN model consist of input node features $X \in R^{n \times F}$ and the adjacency matric $A \in R^{n \times n}$ of the network. Each input node features consist of infectivity, susceptibility, their node labels to encode different inter-contact time between them, and the initial infection information. We can say that the whole network where the spreading process is taking place becomes the input.

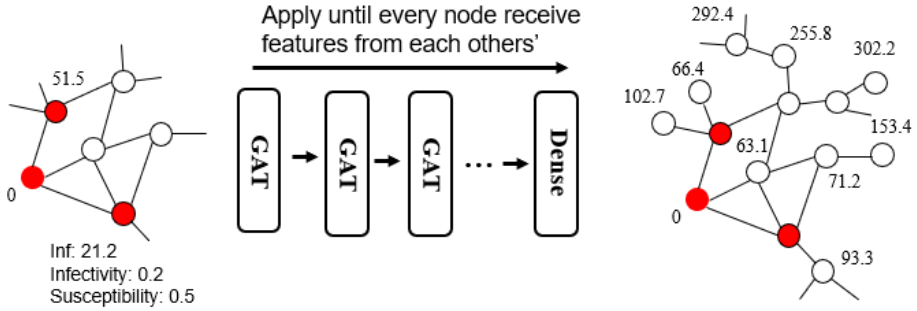

Fig. 3 GAT Model Architecture for Epidemic Prediction

Input node features are propagated and aggregated between adjacent nodes through iterative Message Passing until the readout. Depth of our GNN models are chosen so that by the final iteration of Message Passing every node would have received the node features of every other nodes in the graph at least once to incorporate as many propagation paths as possible.

To predict the local completion time, we need node-level output, i.e, $L(X, v)$ for every node $v \in V$ for which the output of the final Massage Passing is sufficient. To predict the global completion time, we put additional dense layer at the end as readout to output a single value.

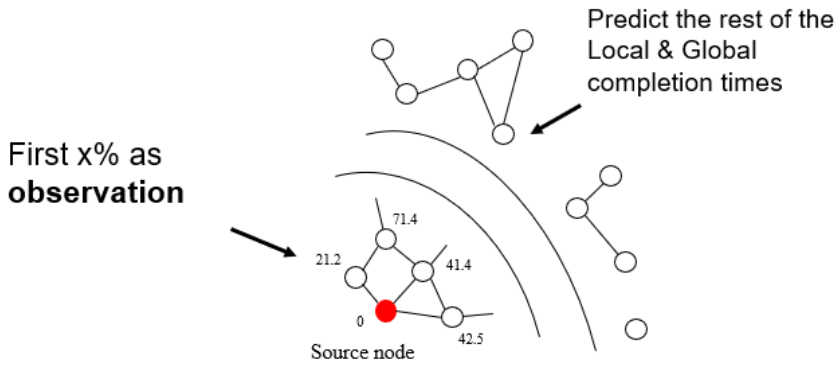

Fig. 4 Encoding Propagation Process into Graphs

The knowledge of a propagation process is provided in the *observations* (Fig. 4) which are the completion time values of the nodes sampled randomly according to chronological order. From the *observations* our GAT models are expected to learn the propagation of information throughout the network eventually to predict the Global and Local Completion time of all the nodes in the network and one value for the entire network.

The provide the potential interpretation, or potentially useful insight, of using the Graph Attention Networks could come from interpretation of attention values. Attentions placed on the neighboring nodes shows which of the adjacent nodes or edges are important when predicting the completion times.

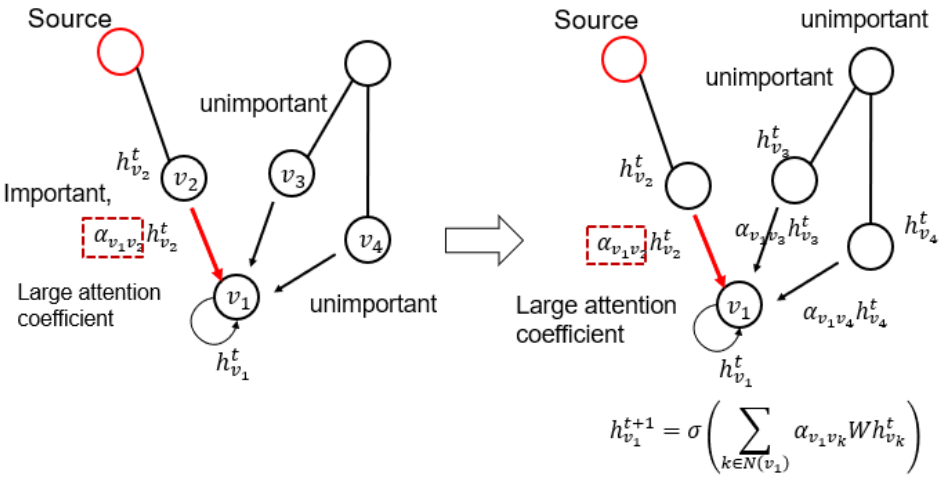

Fig. 5 Interpretation of Attention in Propagation Path of Epidemic models

For example, as in Fig. 5, suppose we are predicting the completion time values for node v_1. Attention values for node v_1 features coming from v_3, v_4 should be very low (those information are not very useful when predicting completion times since there is no possible path that infects v_1 coming from v_3 or v_4. On the contrary, attention values for v_2 should be very high because right behind v_2 there is an adjacent source node which will be very important in deciding the completion time value for node v_1. This interpretation of attention can be of crucial use when deciding for example, links to which nodes should be deleted or added to accelerate or slow down the spread of the information to specific node or graph as whole.

Experimental Evaluation.

We evaluate our proposed epidemic model with GNNs on 10000 networks each with 800 nodes. We first generate networks under three different topologies: classic ones with preferential attachment, and random attachment and practical one from the NY Manhattan island. To simulate spreading processes in a real-world network, we generate the network with the nodes as GPS points on the NY Manhattan island. For the nodes under the Manhattan island topology, we set the mean inter-contact times equal to the mean travel times between the GPS points calculated using Graph Hopper route planning API.

Table 1: Relative Errors (%) of Local Completion Time under Random attachment Topology

Contact Process	1 hop	2 hops	3 hops	4 hops
Shape param. 0.5	12.2803	13.2672	15.5430	16.1547
Shape param. 1.0	7.2754	7.0066	7.5222	9.4668
Shape param. 3.0	5.4753	5.1891	5.3473	5.0251

Table 2: Relative Errors (%) of Local Completion Time under Preferential Attachment Topology

Contact Process	1 hop	2 hops	3 hops
Shape param. 0.5	22.3262	23.9505	22.4616
Shape param. 1.0	15.5117	12.7725	14.6502
Shape param. 3.0	7.1414	5.8177	8.0271

Table 3: Relative Errors (%) of Local Completion Time under Manhattan island Topology

Contact Process	1 hop	2 hops	3 hops
Shape param. 0.5	22.7045	18.4574	11.9917
Shape param. 1.0	11.9902	9.8202	10.4233
Shape param. 3.0	7.0033	5.0134	4.6219

We next diversify the contact patterns between nodes in a way where the inter-contact times follow Weibull distributions with different shape parameters $\beta \in \{0.5, 1.0, 3.0\}$. Note that the shape parameter β characterizes heavytailness of the inter-contact times: β of 0.5, 1.0 and 3.0 corresponds to heavy-tailed, exponential and light-tailed distributions, respectively.

Whenever two nodes contact, there can occur an infection whose possibility is determined by the infectivity and susceptibility of these contacting nodes. We compose each network with heterogeneous nodes by selecting the infectivity and the susceptibility of the nodes by selecting the infectivity and the susceptibility of the nodes uniformly at random in the set $\{0.2, 0.5, 0.7, 1.0\}$. The epidemics are then propagated until every node in the network becomes infected.

We use relative errors to assess the accuracy of our model, defined as:

$$Relative\ Error = \frac{|prediction - true\ value|}{true\ value} \times 100\ (\%)$$

Our model achieves satisfactory performance in predicting both the global and the local completion times. As shown in Figures 6,7 and 8, In predicting the global completion times, our model gives highly accurate predictions with the median relative errors below 15% for every combination of topology, inter-contact time distribution, and *Y*(fraction of the infected nodes in the network).

In Tables 1,2 and 3 we scrutinize the errors involved in predicting the local completion times with respect to the nodes' shortest distance to the closest seed. We can see that the prediction brings a stable error over all hops, meaning that our node-level prediction is impervious to the distance of the node from the seeds.

Relative Errors of Global Completion Time Predictions

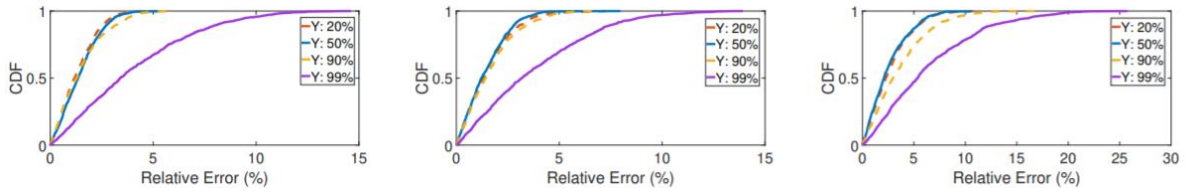

Fig. 6 Global Completion Time under Random Topology (Weibull 0.5, 1.0, 3.0 left to right)

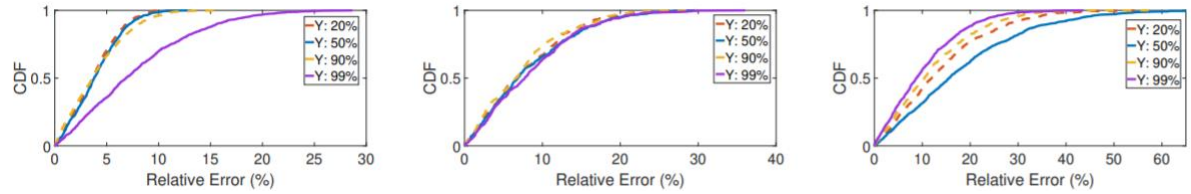

Fig. 7 Global Completion Time under Manhattan Topology (Weibull 0.5, 1.0, 3.0 left to right)

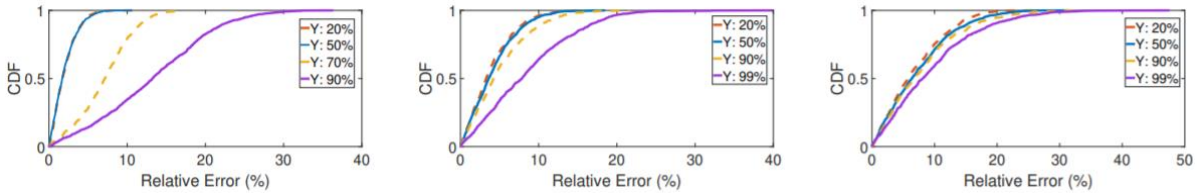

Fig. 8 Global Completion Time under Preferential Topology (Weibull 0.5, 1.0, 3.0 left to right)

Discussion and Conclusion

In this work we developed a new modeling approach that augments conventional epidemic models using GNNs to lift restrictions on contact patterns and network topologies of previous models. Based on our proposed model, we analyze spread times form node-to-network level perspectives in terms on the local and global completion times. Simulation studies show that our proposed approach can predict the spread time with high accuracy for various contact patterns and network topologies with heterogenous nodes.

References

[1] Yoora Kim, Kyunghan Lee, and Ness B Shroff. “On stochastic confidence of information spread in opportunistic networks” IEEE Transactions on Mobile Computing15.4 (2015),pp. 909–923

[2] Petar Velickovic and Guillem Cucurull et al.“Graph Attention Networks”. 6th International Conference on Learning Representations, ICLR 2018

[3] Gilmer, Justin, et al. "Neural message passing for quantum chemistry." *arXiv preprint arXiv:1704.01212* (2017)

[4] Wu, Zonghan, et al. "A comprehensive survey on graph neural networks." *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[5] Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." *arXiv preprint arXiv:1812.08434* (2018).

[6] Rusek, Krzysztof, et al. " Unveiling the potential of Graph Neural Networks for network modeling and optimization in SDN. " *Proceedings of the 2019 ACM Symposium on SDN Research*. 2019.

[7] Sellke, Sarah H., et al. "Modeling and automated containment of worms." *IEEE Transactions on Dependable and Secure Computing* 5.2 (2008): 71-86.

[8] Zhang, Xiaolan, et al. "Performance modeling of epidemic routing." *Computer Networks* 51.10 (2007): 2867-2891.

[9] Koprulu, Irem, Yoora Kim, and Ness B. Shroff. "Battle of opinions over evolving social networks." *IEEE/ACM Transactions on Networking* 27.2 (2019): 532-545.

[10] Zou, Cliff Changchun, Weibo Gong, and Don Towsley. "Code red worm propagation modeling and analysis." *Proceedings of the ACM conference on Computer and communications security*. 2002.